

# Coding or Not: Optimal Mobile Data Offloading in Opportunistic Vehicular Networks

Yong Li, *Member, IEEE*, Depeng Jin, *Member, IEEE*, Zhaocheng Wang, *Senior Member, IEEE*, Lieguang Zeng, and Sheng Chen, *Fellow, IEEE*

**Abstract**—To cope with explosive vehicular traffic and ever-increasing application demands in the vehicular cellular network, opportunistic vehicular networks are used to disseminate mobile data by high-capacity device-to-device communication, which offloads significant traffic from the cellular network. In the current opportunistic vehicular data transmission, coding-based schemes are proposed to address the challenge of opportunistic contact. However, whether coding techniques can be beneficial in the context of vehicular mobile data offloading is still an open question. In this paper, we establish a mathematical framework to study the problem of coding-based mobile data offloading under realistic network assumptions, where 1) mobile data items are heterogeneous in terms of size; 2) mobile users have different interests to different data; and 3) the storage of offloading participants is limited. We formulate the problem as a users' interest satisfaction maximization problem with multiple linear constraints of limited storage. Then, we propose an efficient scheme to solve the problem, by providing a solution that decides when the coding should be used and how to allocate the network resources in terms of contact rate and offloading helpers' storage. Finally, we show the effectiveness of our algorithm through extensive simulations using two real vehicular traces.

**Index Terms**—Erasure coding, mobile data offloading, opportunistic vehicular networks.

## I. INTRODUCTION

WITH THE ever-increasing number of vehicles on roads, traffic jams and accidents have become a serious and widespread problem [1]. Partly to alleviate this serious problem, recently, there has been a strong interest in developing vehicular networks that enable wireless communications for vehicles to obtain information and content from the resource located Internet. Currently, the shortcut for providing such communication

is through cellular networks, such as 3G and 4G. With an increasing number of vehicles equipped with devices to provide vehicular communication capacities through cellular networks, large-scale vehicular cellular access is expected to be available. Many related applications will then emerge, which include safety applications, such as automatic collision warning, remote vehicle diagnostics, emergency management, and assistance for safe driving, as well as other applications, such as automobile high-speed Internet access, vehicle tracking, and multimedia content sharing. Mobile Internet access is becoming increasingly popular for providing various services and applications, including video, audio, and images. According to the latest Cisco forecasts [2], global mobile traffic will increase 18-fold between 2011 and 2016, and monthly global mobile data traffic will surpass 10 EB in 2016. Moreover, two-thirds of the world's mobile data traffic will be video by 2016 [2]. Mobile cellular networks provide the most popular method of mobile access today. With the increase in mobile services and user demands, however, cellular networks will very likely be overloaded and congested in the near future. In particular, during peak time and in urban areas, vehicular communication will face extreme performance hits, in terms of limited network bandwidth, missed calls, and unreliable coverage.

To cope with this explosive growth in traffic demands, it is an urgent agenda for cellular providers to provide quick and promising solutions in vehicular networks. The most straightforward solution is to increase the cellular network capacity by adding more base stations with smaller cell size, such as picocells and femtocells, or by upgrading the cellular network to the next-generation advanced network, such as 4G [3]. However, these activities are financially very expensive with low returns, particularly under the current flat pricing model in which charges are independent of traffic. Even if the capacity of the networks is enhanced this way, the future demands from users and applications will quickly outstretch the network capacity. Some providers are forced to adopt certain short-term methods, e.g., limiting users' traffic to 5 GB per month or educating users on responsible access. Obviously, all the aforementioned methods are ineffective and insufficient. In the long term, providers have to find different network technologies to offer sufficient bandwidth to end users.

From the development of a vehicular network front, the Federal Communications Commission has allocated 75 MHz of spectrum for dedicated short-range communications in vehicular networks [4], and IEEE is also working on standard specifications for vehicular ad hoc networks. Vehicular-to-vehicular communication will be enabled in the near future.

Manuscript received May 2, 2013; revised July 21, 2013; accepted August 13, 2013. Date of publication September 27, 2013; date of current version January 31, 2014. This work is supported by National Basic Research Program of China (Grant 2013CB329001), National Nature Science Foundation of China (Grants 61301080, 61171065, and 61021001), National High Technology Research and Development Program (Grants 2013AA010601 and 2013AA010605), and China's Next Generation Internet (Grant CNGI-12-03-007). The Associate Editor for this paper was G. Yan.

Y. Li, D. Jin, Z. Wang, and L. Zeng are with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: liyong07@tsinghua.edu.cn; jindp@tsinghua.edu.cn; zcwang@tsinghua.edu.cn; zenglg@tsinghua.edu.cn).

S. Chen is with the Department of Electronics and Computer Science, Faculty of Physical and Applied Sciences, University of Southampton, Southampton SO17 1BJ, U.K., and also with the Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia (e-mail: sqc@ecs.soton.ac.uk).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TITS.2013.2281104

Consequently, vehicular mobile data offloading through vehicular ad hoc networks is possible. Since a vehicular network is highly mobile and sometimes sparse, it is hard to maintain a connected network to transmit the traffic [5]. However, opportunistic contacts between vehicles offer higher bandwidth communication capacity for data transmission, which can be utilized to form what is known as the opportunistic vehicular network or Vehicular Delay-Tolerant Network (VDTN) [5]. By exploiting the delay-tolerant nature of non-real-time applications, service providers can delay and even shift the data transmission to VDTN. Benefiting from common interests among the users, providers only need to deliver the information to a small fraction of users, which can then be further disseminated by these selected users through VDTN communications to other users. This kind of traffic offloading is a promising solution to partially support the exponential growth of mobile data [12], which otherwise could not be supported even by the 4G cellular network infrastructure [2].

In the data forwarding of VDTN, the challenge of opportunistic contacts controlled by the unstable network topology has been studied, which is usually addressed by exploiting the node mobility model [4], [6] or using the social network theory [7]. Existing works, such as [6], cope with this challenge by modeling the contact rate and using contact history to predict the communication opportunity. On the other hand, facing the challenge of opportunistic contact in VDTN mobile data offloading, another attractive approach is to use replication (redundancy) and to send identical copies of data simultaneously over multiple paths to mitigate the effects of opportunistic contact [8], [9]. Indeed, erasure coding techniques have been long used to cope with opportunistic and limited transmission capacity [10], which encode a data item into multiple coded packets to provide redundancy. A natural question to investigate is whether erasure coding techniques can be beneficial in the context of VDTN mobile data offloading. Applying erasure coding turns out to be an interesting but nontrivial problem for the following reasons. The first fundamental issue is to determine whether using erasure coding can enhance the efficiency of mobile data offloading. More specifically, we need to know under what network conditions erasure coding is beneficial. We note that erasure coding codes the data into small segments to cope with network opportunistic contact by providing data redundancy. In this way, however, more transmission resources, in terms of buffer space and communication contacts, are needed to transmit the same data, compared with those that do not use erasure coding. Therefore, when the network resource is limited, erasure coding may not be effective in enhancing the system performance. The second issue is that, in the mobile offloading scenario, the buffer to store the mobile data for offloading often has a limited size and, moreover, different nodes have different buffer sizes. Finally, mobile data have different sizes, which adds another dimension to the problem.

In this paper, we investigate the problem of optimal resource allocation with the choice of erasure coding for vehicular-based mobile data offloading. To model a realistic network environment, we consider the following network settings: 1) the network contains heterogeneous vehicles, in terms of network function and data preference; 2) the data items are multitypes

of different sizes; and 3) in the mobile data offloading, the participants' storages are limited and have different sizes. These realistic conditions were not taken into account in the previous works [11]–[13] for simplicity reasons. Our novel contribution is threefold, which can be summarized as follows.

- We formulate the erasure-coding-based optimal mobile data offloading problem with heterogeneous data items and vehicles of limited storages as a utility maximization problem under linear constraints, and further specify the problem into an analyzable form.
- We provide the answer to the question when erasure coding can be used to enhance the system performance. By proposing an efficient algorithm to allocate network resources to mobile data that are waiting for offloading, we provide the decision on which content should or should not use the erasure coding, as well as the coding policy for each mobile data. Based on this result, we propose a data replication algorithm to distribute the coding packets or mobile data into the buffers of offloading helpers.
- Through extensive real trace-driven simulations, we show that our algorithm achieves good system performance in challenging opportunistic vehicular network environments.

The rest of the paper is organized as follows: After presenting the related works in Section II, we describe the system overview and formulate the associated optimization problem in Section III. In Section IV, we analyze the problem and design the related algorithm to obtain the system solution. In Section V, we introduce the experimental environment for performance evaluation and provide simulation results. Finally, we conclude the paper in Section VI.

## II. RELATED WORKS

Recently, the mobile data offloading problem has been investigated under the general context of offloading mobile data from the overloaded cellular networks to other networks [3], [12], [14]–[17]. In this paper, we focus on transmitting the traffic by opportunistic communications between vehicles, i.e., the opportunistic vehicular network offloading. Storage allocation problems are also addressed in the area of traffic-model-based data forwarding and data replication in VDTN [4], [6], [18]–[20]. However, there are significant differences between the erasure-coding-based offloading and the content sharing and routing. The main difference is that our scheme considers how and when to use the coding technologies to enhance the system performance. The other differences are analyzed as follows. First, in the application of mobile data offloading, mobile data originates from the Internet, and most data consist of large files with very different sizes. Content size does not matter in VDTN content sharing, but it has a big impact on mobile data offloading. In our work, we explicitly consider different data sizes and storage constraints. Second, in mobile data offloading, the latency of data matters, since it impacts the user experience. Third, in an offloading problem, the system is more concerned with how much data is offloaded from the cellular network and how much capacity can be saved. Finally, in offloading, nodes are usually mobile users, which can use, for example,

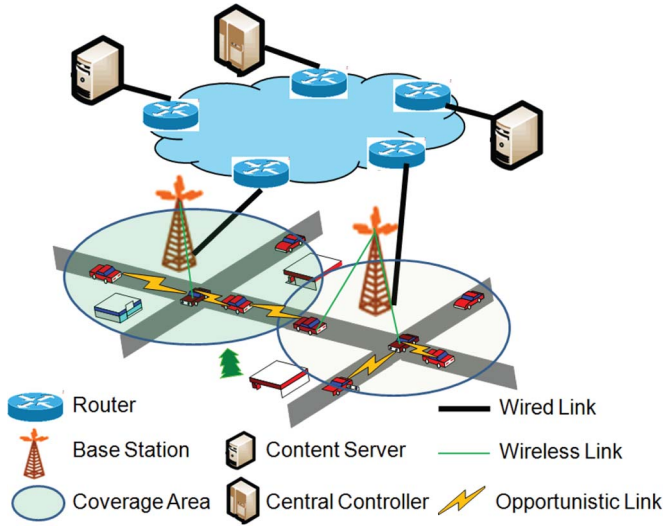


Fig. 1. Illustration of mobile data offloading system integrating cellular network and opportunistic communications.

the 3G network for communication, and the control channel of the cellular system can collect and obtain the overall system metrics of node contact rates and other information. Therefore, centralized algorithms work in this case, and it is not necessary to design distributed algorithms. By contrast, in VDTN content sharing, distribution algorithms are usually required. In addition to these differences with the content sharing problem, our work on mobile data offloading is the first one to investigate the buffer allocation problem with the consideration of coding.

Previous studies have proposed to use network coding to cope with network disruptions in VDTN [29]–[36]. Among them, [32] and [33] provide the erasure coding scheme to increase the network performance. In these works, coding technology was introduced to enhance the system performance of content sharing and dissemination [32]–[36]. However, all these works focus on designing the coding-based routing, forwarding, and data replication schemes by simply assuming that coding would always enhance the system performance. On the other hand, our work focuses on the left fundamental problem that whether coding techniques can be always beneficial in VDTN and provides the solution that decides when and how we could use it to enhance the system performance in the scenario of VDTN mobile data offloading.

### III. SYSTEM OVERVIEW AND PROBLEM FORMULATION

#### A. System Overview

The network topology is shown in Fig. 1, where vehicles travel around the city roads, and a cellular network provides a seamless coverage over the region. All the vehicle stations are dual mode, i.e., they can connect to the cellular network, and they also form the VDTN via opportunistic communications. The base stations of the cellular network are connected to the content servers in the Internet through wireline links. A central controller is deployed for the integrated mobile data offloading system, and it can communicate with the cellular network, vehicle stations, and content servers. The central controller is

tasked to make the coding and data replication decisions for the content servers to distribute the packets of data into the buffers of selected vehicles, based on the vehicular mobility patterns and mobile data demands. Vehicle stations requiring mobile data send their data requests to the corresponding content servers via the cellular network. Then, the requested data are first delivered from the corresponding content servers to some chosen or targeted vehicles via the wireless links of the cellular network under the guidance of the coding and data replication policies. The targeted vehicles will further disseminate the packets of mobile data to the corresponding vehicles that request them through opportunistic communication, which occurs when two vehicles move into the communication range of each other. After the vehicle stations receive the requested data, as usual, they will send acknowledgements to the content servers via the cellular network. Therefore, the content servers know which vehicles have not received their requested data. After the deadline of delivering the mobile data, a content server can directly send the requested data to those vehicle stations that have not received them yet.

#### B. Data Offloading and Networking Modeling

In the aforementioned VDTN-based mobile data offloading system, some chosen vehicles or roadside communication equipment, which are referred to as *helpers*, will participate in the offloading. Incentives for these users can be given by using some micropayment scheme, or the operator can offer the participants a reduced cost for the service or better quality of service (QoS) [21], [22]. A full analysis of such incentives is beyond the scope of this paper. Basically, the service provider chooses some users that are willing to participate in data offloading and transmits mobile data to the chosen users through the cellular network, and then, these users further propagate the data to other users that are interested in them by short-range device-to-device communication. If some users still have not received the data from the helpers after a “tolerable” duration, which is related to the data lifetime, they will directly request and receive the data from the cellular network. More specifically, Fig. 2 illustrates our coding-aided mobile data offloading scheme. There are two types of vehicular nodes in the system, which are known as offloading *helpers* and mobile data *subscribers*, respectively. Mobile data items are first encoded into small segments as packets by erasure coding [10]. Then, the encoded packets are transmitted to the helpers, where they are buffered in the helpers’ local storages. Finally, when the helpers meet other vehicles, they transmit the data to some of them who are interested in these data items. In our system, we consider multiple data offloading, where the system disseminates multiple data items, and a helper may need to store more than one packet, depending on its buffer size. Furthermore, a user may also be interested in different data items.

In our system, there are  $S + H$  mobile users, given by  $i \in \{1, 2, \dots, S + H\}$ . These mobile users are moving vehicles or roadside equipment. Since in reality there are many different types of mobile data, for example, multimedia newspapers, weather forecasts, movie trailers, etc., we model the mobile



traffic of  $C$  different data items, contained in the set  $\mathcal{C}$ , where  $|\mathcal{C}| = C$ . For any  $c \in \mathcal{C}$ , its data length is  $l_c$ , and the lifetime  $T$  is assigned, which means that all the users will not transmit the mobile data  $c$  after the deadline  $T$ . Users, even when some of them have additional storage and energy to help with the traffic offloading, may not be willing to act as helpers due to their selfish behaviors or privacy concerns. Therefore, we use  $\mathcal{H}$  to denote the set of helper users that are willing to participate in the offloading, while using  $\mathcal{S}$  to denote the set of the other subscriber users, where  $|\mathcal{H}| = H$  and  $|\mathcal{S}| = S$ . Note that, from the service provider's perspective, even if many users want to be helpers, it can only choose a limited number of them as helpers as the provider needs to pay directly or indirectly to these users.

If a user acts as helper to participate in the data offloading, the hardware and system resources needed in its mobile devices are the storages to buffer the mobile data, the short-range wireless communication interfaces similar to Bluetooth or WiFi to disseminate the data to other vehicles, and the energy to buffer and transmit the data. In terms of the buffered mobile data, they may include multimedia content of very large size, such as movies. Even the 32 GB of storage available on current devices can only store a limited amount of video content. Furthermore, it is impossible for a user to contribute all its storage for offloading. Therefore, we should take the storage that each helper is willing to share as our constraint, which directly influences the number of data items that can be stored. Considering this realistic condition, we assume that helper  $h$ , i.e.,  $h \in \mathcal{H}$ , can at most buffer  $L_h$  size of data items.

Since we use vehicle-to-vehicle communication to offload data to subscribers, nodes can only communicate when they move into transmission range of each other. Thus, vehicles can communicate with each other only when they move to within the transmission range, which is referred to as a *communication contact*. The contact-based mobility model, which depends of the node mobility velocity and direction, is widely used to model the large-scale vehicular mobility in VDTN [5]–[7]. For the contact-based mobility model, we need to know the patterns of the opportunistic communication contact between vehicles in the mobile data offloading system. In this paper, we assume that communication contact between vehicles  $i$  and  $j$  obeys the Poisson process with contact rate  $\gamma_{i,j}$ . Poisson-distributed contact rate has been validated to fit well to real vehicular traces and is widely used to model opportunistic vehicular systems [23], [24]. The mobility of a helper has much more significant influence on the mobile data offloading system considered than the mobility of a subscriber, and different helpers have different mobility patterns. Therefore, we define the contact rate of helper  $h \in \mathcal{H}$  by  $\gamma_h$ .<sup>1</sup>

In the system, we adopt the erasure coding technique [10], [26] to encode the mobile data into a large set of small coded packets, and any sufficient subset of the coded packets can be used to reconstruct the mobile data. Specifically, the coding process takes the original data  $c$  of size  $l_c$  and a given coding rate as the input and then outputs uniformly sized packets

of size  $s_p$ . Any  $(1 + \epsilon)l_c/s_p$  coded packets can be used to reconstruct the mobile data  $c$  back, where  $\epsilon$  is a small constant determined by the exact erasure coding algorithm employed [26]. In general,  $s_p < l_c$ . Occasionally, mobile data  $c$  may have a very large size, and in this case,  $s_p \ll l_c$ . Without loss of generality, we assume that the minimum coded packet size is unity.<sup>2</sup>

### C. Interest Modeling

We now characterize the subscriber behaviors in accessing to different data items in a mobile data offloading system. In a system with multiple data items, a subscriber will have different interests in different data items, and different subscribers will have different dynamic behaviors. Moreover, some data items are popular data that are interested by many subscribers, whereas some other data items are not popular data, which may only be interesting to a small number of subscribers. In this paper, we describe the subscriber's interests to different mobile data by a subscriber profile and model the popularity of mobile data by an interest distribution. Specifically, for all the mobile data, the system have  $K$  keywords, denoted by the set  $\mathcal{K}$ , to describe them. Any data item  $c \in \mathcal{C}$  is described by a subset of keywords, which is denoted by  $\mathcal{K}_c \subseteq \mathcal{K}$ , and weight  $v_{k_c}$ , which indicates the importance of keyword  $k_c \in \mathcal{K}_c$ . In this way, we can define the popularity of mobile data items. Without loss of generality, we assume  $\sum_{k_c \in \mathcal{K}_c} v_{k_c} = 1$ . To model the interests of different subscribers on different data, we define  $P_s^k$  as the degree of how subscriber  $s \in \mathcal{S}$  is interested in keyword  $k \in \mathcal{K}$ . In this way, we can compare the interests of subscriber  $s$  to two different keywords  $k_1, k_2 \in \mathcal{K}$  by  $P_s^{k_1}$  and  $P_s^{k_2}$ . Thus, the interest profile of subscriber  $s$  is defined by the set  $\mathcal{P}_s = \{P_s^k : k \in \mathcal{K}\}$ . Without loss of generality, we assume  $\sum_{k \in \mathcal{K}} P_s^k = 1$ . Finally, the interest probability of subscriber  $s \in \mathcal{S}$  in mobile data  $c \in \mathcal{C}$ , which is defined by  $w_{s,c}$ , can be obtained as

$$w_{s,c} = \sum_{k_c \in \mathcal{K}_c} v_{k_c} P_s^{k_c}. \quad (1)$$

### D. Problem Formulation

For the mobile offloading system with  $H$  helpers,  $S$  data subscribers, and  $C$  mobile data items, the network optimization goal is to maximize the expected interest satisfaction of all the subscribers in this VDTN system, which depends on the mobile data coding policy and the buffer allocation policy. More specifically, two decisions are required for each mobile data. The first decision is whether erasure coding should be adopted to code the data into small packets, and the other decision is how we replicate the original data or coded packets into the helpers by appropriately allocating their storages. In order to formulate this problem, we first define the coding policy “matrix”  $\mathbf{Y} = (y_c)$ , where  $y_c \in \{1, 2, \dots, l_c - 1, l_c\}$ , and  $y_c$  is the length of the packets that mobile data  $c$  is coded

<sup>1</sup>The assumption is that the contact rate of  $h \in \mathcal{H}$  and any subscriber is the same:  $\gamma_{h,s} = \gamma_h, \forall s \in \mathcal{S}$ . Alternatively,  $\gamma_h$  may be defined as the average contact rate of helper  $h \in \mathcal{H}$  with the subscriber set  $\mathcal{S}$ :  $\gamma_h = E_{s \in \mathcal{S}}[\gamma_{h,s}]$ .

<sup>2</sup>This is effectively using the minimum packet size as the unit for measuring the length of data and buffer.

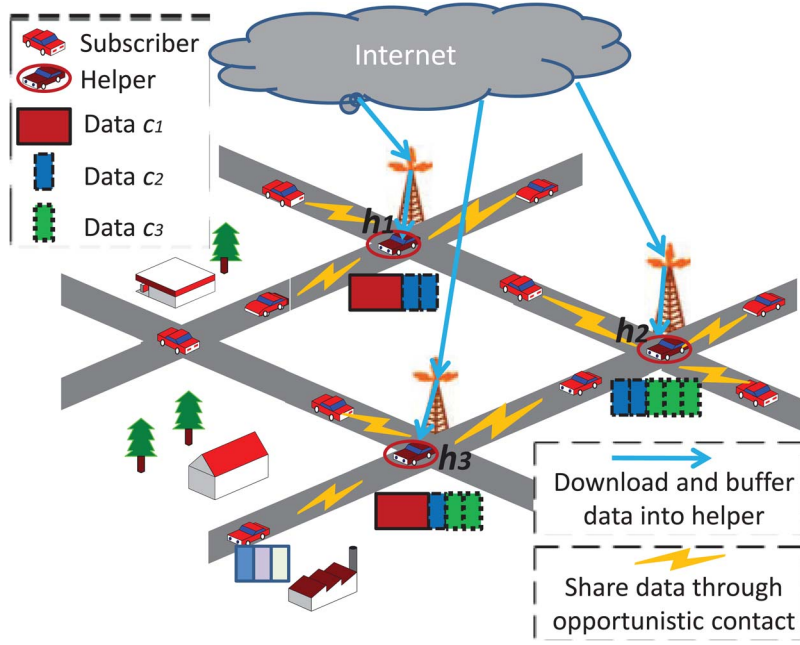


Fig. 2. Illustration of mobile data offloading in opportunistic vehicular networks.

or the length of  $c$  if uncoded. Specifically, we use  $y_c = l_c$  to indicate that mobile data  $c$  is not coded, whereas  $y_c = s_p$  for  $1 \leq s_p < l_c$  indicates that mobile data  $c$  is coded and the length of the coded packets is  $s_p$ . We next define the buffer allocation policy “matrix”  $\mathbf{X}$  to represent the mobile data or packets replication solution in helpers, which is denoted by  $\mathbf{X} = (x_{h,c})$  for  $h \in \mathcal{H}$  and  $c \in \mathcal{C}$  with  $x_{h,c} \in \{0, 1, \dots, l_c - 1, l_c\}$ , where  $x_{h,c}$  is the storage used by helper  $h$  to buffer data  $c$ . More specifically,  $x_{h,c} = 0$  indicates that helper  $h$  does not store the uncoded data item  $c$  or coded packets of  $c$  in its buffer, and  $x_{h,c} = l_c$  means that helper  $h$  stores the whole uncoded data  $c$  in its buffer, whereas  $x_{h,c} = l_{h,c}$ ,  $1 \leq l_{h,c} < l_c$  indicates that helper  $h$  uses the storage of size  $l_{h,c}$  in its buffer to store the coded packets. In the last case, the number of coded packets that can be stored in the buffer of helper  $h$  is  $l_{h,c}/y_c$ .

Taking Fig. 2 for example. In this system, there are three data ( $c_1$ ,  $c_2$ , and  $c_3$ ) and three helpers ( $h_1$ ,  $h_2$ , and  $h_3$ ). For the data coding policy,  $c_1$  is not coded with length of  $l_{c_1}$ , whereas  $c_2$  and  $c_3$  are coded with the data length of 1. Therefore, the coding policy “matrix” for the data is expressed by  $\mathbf{Y} = (l_{c_1}, 1, 1)$ . For the buffer allocation policy,  $h_1$  stores data  $c_1$  and two coded packets of  $c_2$ ;  $h_2$  stores data  $c_1$ , one coded packet of  $c_2$ , and two coded packets of  $c_3$ ; and  $h_3$  stores two coded packets of  $c_2$  and three coded packets of  $c_3$ . Therefore, the buffer allocation policy “matrix”  $\mathbf{X}$  here can be expressed as

$$\mathbf{X} = \begin{pmatrix} l_{c_1} & 2 & 0 \\ l_{c_1} & 1 & 2 \\ 0 & 2 & 3 \end{pmatrix}.$$

With the introduction of  $\mathbf{Y}$  and  $\mathbf{X}$ , i.e., the coding policy and the buffer allocation policy, we may define the system expected interest satisfaction as the objective function  $U(\mathbf{X}, \mathbf{Y})$

and formulate the optimization problem of maximizing the user interest satisfaction as

$$\begin{aligned} \max \quad & U(\mathbf{X}, \mathbf{Y}) \\ \text{s.t.} \quad & y_c \in \{1, 2, \dots, l_c\} \quad \forall c \in \mathcal{C}; \\ & x_{h,c} \in \{0, 1, \dots, l_c\} \quad \forall h \in \mathcal{H}, \quad c \in \mathcal{C}; \\ & \sum_{c \in \mathcal{C}} x_{h,c} \leq L_h \quad \forall h \in \mathcal{H}. \end{aligned} \quad (2)$$

#### IV. PROBLEM ANALYSIS AND ALGORITHM DESIGN

In order to solve the problem (2), we need to obtain the expression of  $U(\mathbf{X}, \mathbf{Y})$ . Since  $U(\mathbf{X}, \mathbf{Y})$  depends on the system dynamics of node contact, data coding policy, and buffer allocation, it is difficult to obtain the expression of  $U(\mathbf{X}, \mathbf{Y})$  explicitly at a glance.

##### A. Problem Specification

However, we note that, when the coding policy is made, i.e., when the decision on whether to code each  $c$  is made and packet size  $s_p$  is chosen for coding data, then  $\mathbf{Y}$  is fixed. In this case, we may express the objective function  $U(\mathbf{X}, \mathbf{Y})$  given  $\mathbf{Y}$  as  $\bar{U}(\mathbf{X})$ , since the system performance is now mainly determined by the system dynamics of node contact and buffer allocation policy. Recall that the goal is to maximize the expected interest satisfaction. The objective function  $\bar{U}(\mathbf{X})$  may be expressed as

$$\bar{U}(\mathbf{X}) = \sum_{s \in \mathcal{S}} \sum_{c \in \mathcal{C}} w_{s,c} Q_{s,c} \quad (3)$$

by explicitly combining the definition of subscribers' interests, where  $Q_{s,c}$  is the probability that subscriber  $s$  has successfully

received mobile data  $c$  before its lifetime. In our system, we assume that a helper will contact any other subscriber with the same contact rate. Therefore, we can drop the index  $s$  in  $Q_{s,c}$  and redefine it as  $V(c) = Q_{s,c}$ ,  $\forall s \in \mathcal{S}$ . Then, the objective function (3) becomes

$$\bar{U}(\mathbf{X}) = \sum_{c \in \mathcal{C}} W_c V(c) \quad (4)$$

where  $W_c = \sum_{s \in \mathcal{S}} w_{s,c}$ .

Now, we derive the expression for  $V(c)$ . If a subscriber  $s$  successfully obtains mobile data  $c$  by decoding the coded packets, it should have received at least  $(1 + \epsilon)l_c/s_p$  packets. Based on the definition of  $V(c)$ , we have

$$\begin{aligned} V(c) &= P \{ \text{Subscriber } s \text{ receives } (1 + \epsilon)l_c/s_p \\ &\quad \text{or more packets} \} \\ &= P \left\{ \sum_{h \in \mathcal{H}} x_{h,c} z_h \geq (1 + \epsilon)l_c \right\} \end{aligned} \quad (5)$$

where the binary indicator variable  $z_h$  takes the values of 1 or 0 with probabilities

$$P(z_h = 1) = \Lambda_h = 1 - e^{-\gamma_h T}; \quad P(z_h = 0) = 1 - \Lambda_h = e^{-\gamma_h T}.$$

Obviously,  $z_h = 1$  indicates that subscriber  $s$  encounters helper  $h$  before the data lifetime  $T$ , whereas  $z_h = 0$  means that subscriber  $s$  does not meet helper  $h$ . Thus,  $z_h$  depends on the contact rate  $\gamma_h$ , which considers the contact rate into the objective functions of  $V(c)$  and  $Q_{s,c}$  in (3) and (4), respectively. Then, the average buffer size used by helpers to store data  $c$  can be defined as

$$N_c = \frac{\sum_{h \in \mathcal{H}, x_{h,c} \neq 0} \Lambda_h x_{h,c}}{\sum_{h \in \mathcal{H}, x_{h,c} \neq 0} \Lambda_h} \geq 1. \quad (6)$$

Note that  $N_c/s_p$  is the average packets that a subscriber can obtain by meeting a helper, and

$$\sum_{h \in \mathcal{H}} x_{h,c} z_h \geq (1 + \epsilon)l_c \Leftrightarrow \sum_{h \in \mathcal{H}} \frac{x_{h,c}}{N_c} z_h \geq \frac{(1 + \epsilon)l_c}{N_c} \triangleq m_c. \quad (7)$$

When  $m_c \gg \gamma_h$  and  $(x_{h,c}/N_c) \approx 1$ , we may use the Poisson process to approximate the distribution of  $\sum_{h \in \mathcal{H}} (x_{h,c}/N_c) z_h$ , i.e.,  $\sum_{h \in \mathcal{H}} (x_{h,c}/N_c) z_h$  follows the Poisson distribution with the parameter  $d_c/N_c$ , where

$$d_c = \sum_{h \in \mathcal{H}} \Lambda_h x_{h,c}. \quad (8)$$

Thus, we arrive at

$$\begin{aligned} V(c) &= P \left\{ \sum_{h \in \mathcal{H}} \frac{x_{h,c}}{N_c} z_h \geq m_c \right\} = \Gamma \left( \frac{d_c}{N_c}, m_c \right) \\ &= \Gamma \left( \frac{d_c}{(1 + \epsilon)l_c}, m_c \right) \end{aligned} \quad (9)$$

where  $\Gamma(\alpha, \beta)$  is the Gamma distribution, with the expression of

$$\Gamma(\alpha, \beta) = \frac{\int_0^\alpha e^{-x} x^{\beta-1} dx}{(\beta - 1)!}.$$

In the aforementioned expression of  $V(c)$ , we note that  $d_c$  is the equivalent network resource, which depends on the allocated buffer for mobile data  $c$  and the contact rates of helpers. We will refer to  $d_c$  as the system allocated network resource for mobile data  $c$ . Therefore,  $d_c$ ,  $c \in \mathcal{C}$  are the variables related to  $\mathbf{X}$ . Furthermore, since at least  $(1 + \epsilon)l_c/s_p$  packets are needed to decode data  $c$ , we define the resource abundant rate as

$$\eta_c = \frac{d_c}{(1 + \epsilon)l_c}. \quad (10)$$

In addition,  $m_c$  represents the expected number of helpers a subscriber should contact before it could receive sufficient packets to decode data  $c$ . Two special cases or two extreme values of  $m_c$  are worth mentioning. The case of  $m_c = 1$  indicates that the subscriber only needs to meet one helper, which is related to the scenario that the uncoded data  $c$  is stored in this helper, whereas  $m_c = (1 + \epsilon)l_c$  means that the subscriber needs to meet  $(1 + \epsilon)l_c$  helpers to receive the required number of packets, which is related to the scenario that the data item is encoded by the minimum length of  $s_p = 1$ . Therefore,  $m_c$  is related to the erasure coding policy, and in general  $1 \leq m_c \leq (1 + \epsilon)l_c$ . It turns out that  $m_c$  for  $c \in \mathcal{C}$  represents the coding policy  $\mathbf{Y}$ .

Let  $\mathbf{d}$  and  $\mathbf{m}$  be the vectors that contain all the elements of  $d_c$  and  $m_c$ , for  $c \in \mathcal{C}$ , respectively. We have derived the following expression for the objective function  $U(\mathbf{X}, \mathbf{Y})$ :

$$\begin{aligned} \hat{U}(\mathbf{d}, \mathbf{m}) &= \sum_{c \in \mathcal{C}} W_c V(c) = \sum_{c \in \mathcal{C}} W_c \Gamma \left( \frac{d_c}{(1 + \epsilon)l_c}, m_c \right) \\ &= \sum_{c \in \mathcal{C}} W_c \Gamma(\eta_c m_c, m_c) \end{aligned} \quad (11)$$

and transferred the optimization problem (2) into the following form:

$$\begin{aligned} \max \quad & \hat{U}(\mathbf{d}, \mathbf{m}) \\ \text{s.t.} \quad & d_c \in \mathbb{R}, m_c \in \{1, 2, \dots, (1 + \epsilon)l_c\} \quad \forall c \in \mathcal{C}; \\ & \sum_{c \in \mathcal{C}} d_c = d_{\text{all}} \end{aligned} \quad (12)$$

where the constant

$$\begin{aligned} d_{\text{all}} &= \sum_{c \in \mathcal{C}} \sum_{h \in \mathcal{H}} \Lambda_h x_{h,c} = \sum_{h \in \mathcal{H}} \Lambda_h \sum_{c \in \mathcal{C}} x_{h,c} \\ &= \sum_{h \in \mathcal{H}} (1 - e^{-\gamma_h T}) L_h \end{aligned}$$

is obtained by taking the equality sign in the inequality  $\sum_{c \in \mathcal{C}} x_{h,c} \leq L_h$ . Since  $m_c$  indicates the expected number of packets needed to encode data  $c$ , it may be treated as integer valued to simplify the analysis, as done in forming the optimization problem (12).

### B. Algorithm Design

For the optimization problem (12), we observe that we need to decide how to allocate the system resource  $d_{\text{all}}$  to each mobile data and also need to choose the data coding policy. We have two theorems regarding the properties of the objective function  $\hat{U}(\mathbf{d}, \mathbf{m})$ , with respect to the decision variables  $\mathbf{d}$  and  $\mathbf{m}$ , respectively. First, we introduce the following lemma.

**Lemma 1:** The function

$$J(m_c) = \eta_c \int_0^1 e^{-\eta_c x} \left( \frac{x}{m_c} + 1 \right)^{m_c} dx - 1 \quad (13)$$

is a monotonic increasing function of  $m_c$ .

*Proof:* If we can prove that  $g_x(m_c) = ((x/m_c) + 1)^{m_c}$  is a monotonic increasing function of  $m_c$  given the condition that  $0 \leq x \leq 1$  and  $m_c \geq 1$ , then for any  $1 \leq m_{c1} < m_{c2} < +\infty$ , we have

$$J(m_{c1}) - J(m_{c2}) = \eta_c \int_0^1 e^{-\eta_c x} (g_x(m_{c1}) - g_x(m_{c2})) dx < 0$$

based on which we can conclude that  $J(m_c)$  is a monotonic increasing function of  $m_c$ .

For  $0 \leq x \leq 1$  and  $m_c \geq 1$ ,  $g_x(m_c) > 0$ . Define

$$h(m_c) \triangleq \ln(g_x(m_c)) = m_c \cdot \ln \left( \frac{x}{m_c} + 1 \right).$$

Note

$$h'(m_c) = \ln \left( 1 + \frac{x}{m_c} \right) - \frac{\frac{x}{m_c}}{1 + \frac{x}{m_c}}.$$

By denoting  $a = (x/m_c)$ , we have

$$h'(m_c) = \ln(1+a) - \frac{a}{1+a} \triangleq e(a), \quad 0 < a < 1.$$

Now  $e(a) > 0$  for  $0 < a < 1$ , since  $e(0) = 0$  and  $e'(a) = (a/(1+a)^2) > 0$  for  $0 < a < 1$ . This proves that  $h(m_c)$  is a monotonic increasing function of  $m_c$ , and so is  $g_x(m_c)$ . ■

**Theorem 1:** In order to maximize the objective function (11),  $\forall c \in \mathcal{C}$ ,  $m_c$  should take the value of 1 or  $(1+\epsilon)l_c$ .

*Proof:* To analyze the influence of  $\mathbf{m}$ , we fix  $\mathbf{d}$ . Noting that  $\hat{U}(\mathbf{d}, \mathbf{m})$  is a linear combination of  $V(c)$ , we simply denote  $\hat{V}(m_c) = V(c)$ . As mentioned before,  $m_c$  indicates the expected number of packets needed to encode data  $c$ , and we may assume it is integer valued. Taking the first-order difference of  $\hat{V}(m_c)$  with respect to  $m_c$  leads to

$$\begin{aligned} \Delta \hat{V}(m_c) &= \Delta (W_c \Gamma(\eta_c m_c, m_c)) \\ &= W_c (\Gamma(\eta_c(m_c+1), (m_c+1)) - \Gamma(\eta_c m_c, m_c)) \\ &= W_c \frac{e^{-\eta_c m_c} (\eta_c m_c)^{m_c}}{m_c!} \\ &\quad \times \left( \eta_c \int_0^1 e^{-\eta_c x} \left( \frac{x}{m_c} + 1 \right)^{m_c} dx - 1 \right) \\ &= W_c \frac{e^{-\eta_c m_c} (\eta_c m_c)^{m_c}}{m_c!} J(m_c). \end{aligned}$$

According to Lemma 1,  $J(m_c)$  is a monotonic increasing function of  $m_c$ . Consequently, the monotonicity property of  $V(c)$  belongs to the following three cases: 1) monotonic increasing; 2) monotonic decreasing; and 3) first decreasing and then increasing. Furthermore, since  $1 \leq m_c \leq (1+\epsilon)l_c$ , we have

$$\frac{1 - e^{-\eta_c} - 2\eta_c e^{-\eta_c}}{\eta_c} \leq J(m_c) \leq \frac{\eta_c e^{1-\eta_c} - 1}{1 - \eta_c}.$$

Therefore, we further have the following observations.

- When  $\eta_c < 1$ ,  $J(m_c) < 0$ . Consequently,  $\Delta \hat{V}(m_c) < 0$  and  $\hat{V}(m_c)$  decreases with the increase in  $m_c$ .
- When  $\eta_c > 1.25$ ,  $J(m_c) > 0$ . Consequently,  $\Delta \hat{V}(m_c) > 0$  and  $\hat{V}(m_c)$  increases with the increase in  $m_c$ .
- When  $1 \leq \eta_c \leq 1.25$ , depending on  $\eta_c$ ,  $\hat{V}(m_c)$  may be monotonic increasing, monotonic decreasing, or first decreasing and then increasing.

Combining the previous three cases, we conclude that depending on the system allocated resource to mobile data  $c$ , i.e.,  $\eta_c$ ,  $m_c$  should take the value of 1 or  $(1+\epsilon)l_c$  in order to maximize the objective function  $\hat{U}(\mathbf{d}, \mathbf{m})$ . ■

**Remark 1:** Note that  $m_c = (1+\epsilon)l_c$  corresponds to choosing the minimum packet size of  $s_p = 1$  in erasure coding, whereas  $m_c = 1$  indicates that no erasure coding is applied. According to Theorem 1, for any data  $c$ , if  $\eta_c < 1$ , then the optimal system performance is obtained with  $m_c = 1$ , which means that we should not apply erasure coding to data  $c$ . Rather, the whole data  $c$  should be stored as a packet into the helpers. This makes sense. From the definition of the resource abundant rate given in (10),  $\eta_c < 1$  indicates that the allocated network resource  $d_c$  is definitely insufficient to cope with the requirement of coding data  $c$ . In such a case, it is best not to code  $c$ . By contrast, if  $\eta_c > 1.25$ , which indicates that the allocated network resource is certainly sufficient, then we should apply erasure coding for data  $c$  and encode it into packets with the minimum packet size of  $s_p = 1$ . The situation of  $1 \leq \eta_c \leq 1.25$  is more complicated, but the decision of not coding data  $c$ , or coding it with the minimum packet size of  $s_p = 1$ , can be made once the value of  $\eta_c$  is obtained.

From the aforementioned discussion, we can see that, once the optimal values of  $\eta_c$  (i.e.,  $d_c$ ) for every data  $c$  are obtained, it becomes relatively easy to solve the optimization problem (12). To analyze the influence of  $\mathbf{d}$  on  $\hat{U}(\mathbf{d}, \mathbf{m})$ , we fix  $\mathbf{m}$ . We note that when the objective function is maximized,  $m_c$  takes the value of either 1 or  $(1+\epsilon)l_c$ , according to  $\eta_c$ . Therefore, we define

$$V_1(\eta_c) = \sum_{c \in \mathcal{C}} W_c \Gamma(\eta_c, 1) \quad (14)$$

$$V_2(\eta_c) = \sum_{c \in \mathcal{C}} W_c \Gamma(\eta_c(1+\epsilon)l_c, (1+\epsilon)l_c). \quad (15)$$

We have the following two lemmas for the functions  $\Gamma(\eta_c, 1)$  and  $\Gamma(\eta_c(1+\epsilon)l_c, (1+\epsilon)l_c)$ , respectively.

**Lemma 2:**  $\Gamma(\eta_c, 1)$  is a convex function.



*Proof:* Note that  $\Gamma(\eta_c, 1) = 1 - e^{-\eta_c}$ . Since

$$\frac{\partial^2 \Gamma(\eta_c, 1)}{\partial \eta_c^2} = -e^{-\eta_c} < 0$$

we conclude that  $\Gamma(\eta_c, 1)$  is a convex function. ■

**Lemma 3:**  $\Gamma(\eta_c(1 + \epsilon)l_c, (1 + \epsilon)l_c)$  is first a concave function and then becomes a convex function.

*Proof:* By denoting  $m = (1 + \epsilon)l_c$ , we have

$$\begin{aligned} \Gamma(\eta_c(1 + \epsilon)l_c, (1 + \epsilon)l_c) &= \Gamma(\eta_c m, m) \\ &= \frac{\int_0^{\eta_c m} e^{-x} x^{m-1} dx}{(m-1)!}. \end{aligned}$$

Furthermore

$$\frac{\partial^2 \Gamma(\eta_c m, m)}{\partial \eta_c^2} = \eta_c m^2 e^{-\eta_c m} (\eta_c m)^{m-1} \left( \frac{m-1}{m} - \eta_c \right). \quad (16)$$

The point at which the second derivative becomes zero is known as the *inflection point* of the original function. From (16), it is easy to see that the inflection point of  $\Gamma(\eta_c m, m)$  is

$$\eta_{c, \text{inflection}} = \frac{m-1}{m} \quad (17)$$

which is between 0 and 1. Since the second derivative function (16) turns from positive to negative at the inflection point  $\eta_{c, \text{inflection}}$ , we conclude that  $\Gamma(\eta_c m, m)$  is first a concave function and then becomes a convex function. ■

We now define

$$V_0(\eta_c) = \max \{V_1(\eta_c), V_2(\eta_c)\} \quad (18)$$

The convexity of  $V_0(\eta_c)$  is stated in the following theorem.

**Theorem 2:** There exists a tangent line upper bounding  $V_0(\eta_c)$ , which touches  $V_1(\eta_c)$  and  $V_2(\eta_c)$  at  $\eta_c = \eta_1$  and  $\eta_2$ , respectively, where  $1 < \eta_1 < \eta_2 < 1.25$ . Moreover,  $V_0(\eta_c)$  is a convex function for  $\eta_c \leq \eta_1$  and  $\eta_c \geq \eta_2$ .

*Proof:* From the proof of Theorem 1, it is easy to see that  $V_0(\eta_c) = V_1(\eta_c)$  is convex for  $\eta_c < 1$  and  $V_0(\eta_c) = V_2(\eta_c)$  is convex for  $\eta_c > 1.25$ . Furthermore, according to Lemmas 2 and 3, both  $V_1(\eta_c)$  and  $V_2(\eta_c)$  are convex for  $1 < \eta_c < 1.25$ . It is then obvious that there exists a single point  $1 < \eta_{c0} < 1.25$  such that  $V_1(\eta_{c0}) = V_2(\eta_{c0})$ . Thus,  $V_0(\eta_c) = V_1(\eta_c) > V_2(\eta_c)$  for  $\eta_c < \eta_{c0}$ , and  $V_0(\eta_c) = V_2(\eta_c) > V_1(\eta_c)$  for  $\eta_c > \eta_{c0}$ . In both cases,  $V_0(\eta_c)$  is convex.

Therefore,  $V_0(\eta_c)$  looks like two convex functions joined at  $\eta_c = \eta_{c0}$ . Now step a straight line on the top of  $V_0(\eta_c)$ . With the help of geometric intuition, we can see that this line touches  $V_0(\eta_c)$  only at two points  $\eta_1$  and  $\eta_2$  with  $1 < \eta_1 < \eta_{c0} < \eta_2 < 1.25$  and  $V_1'(\eta_1) = V_2'(\eta_2)$ . ■

The proof of Theorem 2 is illustrated in Fig. 3. We will refer to this upper bound line of  $V_0(\eta_c)$  as its *upper bounding tangent line*.

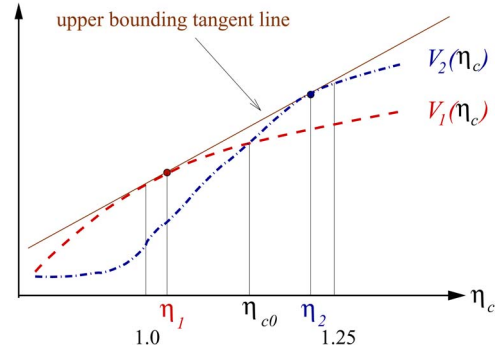


Fig. 3. Illustration of geometric properties of the functions  $V_1(\bullet)$ ,  $V_2(\bullet)$ , and  $V_0(\bullet) = \max\{V_1(\bullet), V_2(\bullet)\}$ .

---

#### Algorithm 1 Resource Allocation.

---

- 1: Define  $\hat{V}(d_c) = V_0(\eta_c)$  for  $\eta_c < \eta_1$  or  $\eta_c > \eta_2$ ; otherwise  $\hat{V}(d_c)$  is set to the upper bounding tangent line of  $V_0(\eta_c)$
  - 2: Initial  $\Phi_{\text{up}} = \max\{W_c\}$ ,  $\Phi_{\text{down}} = 0$ ,  $\Phi_{\text{mid}} = 0$  and precision parameter  $\varpi = 10^{-3}$
  - 3: **while**  $|\Phi_{\text{up}} - \Phi_{\text{down}}| > \varpi$  **do**
  - 4:    $\Phi_{\text{mid}} = (\Phi_{\text{up}} + \Phi_{\text{down}}/2)$ ,  $d_{\text{used}} = 0$
  - 5:   **for**  $c \in \mathcal{C}$  **do**
  - 6:     **if**  $\Phi_{\text{mid}} \leq \hat{V}'(d_c)|_{d_c=0} \cdot W_c$  **then**
  - 7:       Calculate  $d_c$ , s.t.  $\hat{V}'(d_c) \cdot W_c = \Phi_{\text{mid}}$
  - 8:        $d_{\text{used}} = d_{\text{used}} + d_c$
  - 9:     **else**
  - 10:        $d_c = 0$
  - 11:     **end if**
  - 12:   **end for**
  - 13:   **if**  $d_{\text{used}} > d_{\text{all}}$  **then**
  - 14:      $\Phi_{\text{mid}} = \Phi_{\text{up}}$
  - 15:   **else**
  - 16:      $\Phi_{\text{mid}} = \Phi_{\text{down}}$
  - 17:   **end if**
  - 18: **end while**
  - 19: **for**  $c \in \mathcal{C}$  **do**
  - 20:    $\eta_c = (d_c / (1 + \epsilon)l_c)$
  - 21:   **if**  $\eta_c \leq \eta_1$  **then**
  - 22:      $m_c = 1$
  - 23:   **else**
  - 24:      $m_c = (1 + \epsilon)l_c$
  - 25:   **end if**
  - 26: **end for**
- 

*1) Algorithm for Resource Allocation and Coding Policy Selection:* Based on Theorems 1 and 2, we design an optimal algorithm, which is summarized in Algorithm 1, to solve the optimization problem (12). This algorithm first decides the optimal resource allocation policy for  $\mathbf{d}$  and then makes the optimal decision on erasure coding policy, i.e., deciding whether  $m_c$  takes the value of 1 or  $(1 + \epsilon)l_c$ . The optimality of Algorithm 1 is proved in the following theorem.



**Theorem 3:** The solution obtained by Algorithm 1 is the optimal solution for the optimization problem (12).

*Proof:* According to Theorem 2, the function  $\hat{V}(d_c)$  defined in Algorithm 1 is a monotonically increasing and convex function. Thus,  $\hat{V}'(d_c) > 0, \forall d_c \geq 0$ . It is clear that the solution obtained by Algorithm 1 for  $\mathbf{d}$  meets the following conditions:

$$W_c \cdot \hat{V}'(d_c) = C_{A1} \quad \forall d_c \neq 0 \quad (19)$$

$$W_c \cdot \hat{V}'(d_c) < C_{A1} \quad \forall d_c = 0. \quad (20)$$

Here, the constant  $C_{A1}$  is equal to  $\Phi_{\text{mid}}$  obtained in Algorithm 1. Therefore, we need to prove that the objective function

$$\hat{U}(\mathbf{d}, \mathbf{m}) = \sum_{c \in \mathcal{C}} W_c \hat{V}(d_c) \quad (21)$$

achieves the maximum value if and only if the conditions (19) and (20) are met, where  $m_c, \forall c \in \mathcal{C}$ , have been set to their corresponding optimal values of either 1 or  $(1 + \epsilon)l_c$ .

We offer the proof in three steps. First, the range of  $\mathbf{d}$  satisfies the condition

$$d_c \geq 0 \quad \forall c \in \mathcal{C}; \quad \sum_{c \in \mathcal{C}} d_c = d_{\text{all}}. \quad (22)$$

This defines a closed set. Since a continuous function has at least one maximum on a closed set, the maximum value of  $\hat{U}(\mathbf{d}, \mathbf{m})$  does exist.

Second, assume that  $\mathbf{d}$  achieves the maximum value of the objective function (21). Consider  $d_{c1} \leq d_{c2}$ . Let us make the following perturbation:

$$d_{c1}^{\text{pert}} = d_{c1} + \sigma; \quad d_{c2}^{\text{pert}} = d_{c2} - \sigma \quad (23)$$

where  $\sigma$  is infinitesimal. We have

$$\begin{aligned} \hat{U}^{\text{pert}} - \hat{U} &= W_{c1} \left( \hat{V}(d_{c1}^{\text{pert}}) - \hat{V}(d_{c1}) \right) \\ &\quad + W_{c2} \left( \hat{V}(d_{c2}^{\text{pert}}) - \hat{V}(d_{c2}) \right) \\ &= \sigma \left( W_{c1} \cdot \hat{V}'(d_{c1}) - W_{c2} \cdot \hat{V}'(d_{c2}) \right). \end{aligned} \quad (24)$$

Since  $\mathbf{d}$  maximizes  $\hat{U}(\mathbf{d}, \mathbf{m})$ ,  $\hat{U}^{\text{pert}} - \hat{U} \leq 0$ . Note that  $d_c \geq 0, \forall c \in \mathcal{C}$ , and the perturbation (23) must be legitimate (in the range of  $\mathbf{d}$ ). This rules out  $d_{c1} = d_{c2} = 0$ , since  $\sigma$  can be positive or negative. The two legitimate cases are the following: 1)  $d_{c1} > 0$ , and  $\sigma$  is either positive or negative; 2)  $d_{c1} = 0$ ,  $\sigma$  is positive, and  $d_{c2} > 0$ . In the first case,  $\hat{U}^{\text{pert}} - \hat{U} \leq 0$  leads to

$$W_{c1} \cdot \hat{V}'(d_{c1}) = W_{c2} \cdot \hat{V}'(d_{c2}). \quad (25)$$

Thus, we must have  $\forall d_c \neq 0, W_c \cdot \hat{V}'(d_c) = \text{Constant}$ . In the second case, we must have

$$W_{c1} \cdot \hat{V}'(d_{c1}) < W_{c2} \cdot \hat{V}'(d_{c2}) = \text{Constant} \quad (26)$$

that is, we must have  $W_c \cdot \hat{V}'(d_c) < \text{Constant}, \forall d_c = 0$ . Combining both cases proves that the optimal solution of the problem (12) meets the conditions (19) and (20).

Third, we observe that there exists only one point generated by Algorithm 1 that satisfies the conditions (19) and (20). This completes the proof. ■

**Remark 2:** Algorithm 1 obtains all the optimal resource allocation decisions  $\{d_c\}$  and almost all the optimal coding decisions  $\{m_c\}$ ,  $\forall c \in \mathcal{C}$ . This is because the algorithm sets  $m_c = (1 + \epsilon)/l_c$  for  $\eta_c > \eta_1$ . According to Theorem 2, there exists a very small interval  $(\eta_1, \eta_2)$ , where the optimal coding decision can be either  $m_c = 1$  or  $m_c = (1 + \epsilon)/l_c$ , depending on the value of  $\eta_c$  and the specific objective function. In other words, there is a very small probability that, for some  $\eta_c \in (\eta_1, \eta_2)$ , the optimal coding decision is  $m_c = 1$ . Note that, since  $\eta_c > \eta_1 > 1$ , the allocated optimal network resource  $d_c$  is sufficient for coding data  $c$  and, therefore, choosing  $m_c = (1 + \epsilon)/l_c$  will not cause any difficulty.

2) **Data Replication Algorithm:** We now provide an efficient algorithm to replicate the original mobile data or encoded packets into the buffer of each helper. The aim of this algorithm is to obtain  $\{x_{h,c}\}$  based on the allocation constraint of  $\mathbf{m}$  and  $\mathbf{d}$  obtained in Algorithm 1. Since the constraint is loose, a heuristic algorithm works well here.

---

#### Algorithm 2 Heuristic Algorithm for Data Replication.

---

- 1: Define  $C_{nc} = \{c | m_c = 1\}$
  - 2: Sort  $C_{nc}$  by descending order of  $W_c : C_{nc} = \{c_1, c_2, \dots, c_{|C_{nc}|}\}$  with  $W_{c1} \geq W_{c2} \geq \dots \geq W_{c_{|C_{nc}|}}$
  - 3: **for**  $i = 1$  to  $|C_{nc}|$  **do**
  - 4:   Loop for  $\lfloor (d_{c_i}/\Lambda_{\text{ave}} \cdot l_{c_i}) \rfloor$  times, select the maximum  $L_h > 0$ , s.t.  $x_{h,c_i} = 0$
  - 5:   **if** there is no such  $h$  **then**
  - 6:     Break the loop
  - 7:   **end if**
  - 8:   **if**  $L_h \geq l_{c_i}$  **then**
  - 9:      $L_h = L_h - l_{c_i}, x_{h,c_i} = l_{c_i}, d_{c_i} = d_{c_i} - l_{c_i} \cdot \Lambda_h$
  - 10:   **end if**
  - 11: **end for**
  - 12: Define  $C_{uc} = \{c | m_c = (1 + \epsilon)l_c\}$
  - 13: Sort  $C_{uc}$  by descending order of  $W_c : C_{uc} = \{c_1, c_2, \dots, c_{|C_{uc}|}\}$  with  $W_{c1} \geq W_{c2} \geq \dots \geq W_{c_{|C_{uc}|}}, i = 0$
  - 14: **while**  $\max\{L_h\} > 2 \cdot \min\{L_h\}$  **do**
  - 15:    $i = (i \bmod |C_{uc}|) + 1$ , Until  $d_{c_i} \geq \Lambda_{\text{ave}}$
  - 16:   Select the maximum  $L_h > 0$ , s.t.  $x_{h,c_i} < l_{c_i}$
  - 17:    $L_h = L_h - 1, x_{h,c_i} = x_{h,c_i} + 1, d_{c_i} = d_{c_i} - \Lambda_h$
  - 18: **end while**
  - 19:  $i = 0, j = 0$
  - 20: **while**  $\max\{L_h\} \neq 0$  **do**
  - 21:    $j = (j \bmod |\mathcal{H}|) + 1$ , Until  $L_{h_j} \neq 0$
  - 22:    $i = (i \bmod |C_{uc}|) + 1$ , Until  $d_{c_i} \geq \Lambda_{h_j}$
  - 23:   **if** there is no such  $i$  or  $j$  **then**
  - 24:     Break the loop
  - 25:   **end if**
  - 26:    $L_{h_j} = L_{h_j} - 1, x_{h_j,c_i} = x_{h_j,c_i} + 1, d_{c_i} = d_{c_i} - \Lambda_{h_j}$
  - 27: **end while**
  - 28: **if**  $\max\{L_h\} \neq 0$  **then**
  - 29:   Randomly fill the buffer with any data
  - 30: **end if**
-

The designed algorithm is shown in Algorithm 2, where parameter  $\Lambda_{\text{ave}} = (1/H) \sum_{h \in \mathcal{H}} \Lambda_h$ . The algorithm contains three steps. In the first step, we allocate the buffer for uncoded data ( $m_c = 1$ ) by considering each content in decreasing order of  $W_c$ . We use the helper with the largest remaining buffer size to store each content to fulfill the constraint  $\{d_c\}$ . In the second step, we allocate the buffer for coded data ( $m_c = (1 + \epsilon)l_c$ ). We loop the contents by decreasing order of  $W_c$  until all the contents fulfill their allocation constraints and, each time at the loop, only one packet is put into buffer if its allocated network resource has not been reached. The helpers' buffers are also looped by a fixed order. First, we loop the helpers whose remaining buffer sizes are larger than twice the minimum buffer size to make the remaining buffers similar sizes, until there is no such helper. Second, we loop the helpers with decreasing order of remaining buffer size. After these two loops, almost all the data contents can fulfill their allocation constraints. In the third step, if there are still remaining buffers in any helpers, we allocate them randomly.

*Remark 3:* Our proposed erasure-coding-based data offloading scheme is a centralized one, which is run at the central controller. The central controller requires global information of the content lengths  $l_k$ , the storage sizes of helper  $L_h$ , the subscribers' content interests  $w_{s,c}$ , and the contact rates  $\gamma_{i,j}$ . Note that these information are not collected via the opportunistic vehicular network. Since the central controller is connected to the content servers via the wired Internet, it can obtain the content-related parameters of  $l_k$  with ease. On the other hand, vehicles are equipped with communication interfaces to connect to the cellular network. Therefore, they are able to send the vehicle-related parameters of  $L_h$ ,  $w_{s,c}$ , and  $\gamma_{i,j}$  through the uplink channel of the cellular network to the central controller. On the other hand,  $l_k$ ,  $L_h$ , and  $w_{s,c}$  are statistic parameters, which are well known by the content servers or the central controller in advance. Many vehicles travel on predetermined routes and schedules, and examples include city buses and people traveling by cars to and from work. Therefore, daily mobility patterns exhibit certain regularity, and the contact rates between many vehicles are often quasi-statistic, which can often be obtained by the central controller in advance with high accuracy.

We can now summarize this section. Using Algorithm 1, we obtain the erasure coding policy  $\{m_c\}$  and the network resource allocation policy  $\{d_c\}$  for each content  $c \in \mathcal{C}$  to solve the optimization problem defined in (12). Then, with Algorithm 2, we obtain the buffer allocation scheme  $x_{h,c}$ ,  $\forall c \in \mathcal{C}$  and  $\forall h \in \mathcal{H}$ . Thus, our solution offers an optimal mobile data offloading implementation for realistic VDTNs, in terms of maximizing the users' interest satisfaction. In the mobile data offloading system, the proposed Algorithms 1 and 2 are running on the central controller, which is usually located in a deployed server in the Internet. Consequently, it should be powerful for the computation, particularly with the current extensive usages of the cloud computing. On the other hand, both Algorithms 1 and 2 have pseudo-polynomial-time complexity, which is acceptable for common servers because of their high computational capacity, and usually in real-world networks, the number of helpers is limited and will not contribute much to the complex-

ity. Therefore, there is no problem for the central controller to work in real time in a practical system.

## V. PERFORMANCE EVALUATION

In a most generic VDTN-based content dissemination system, the most important performance metrics include the fraction of the offloaded contents and the content dissemination delay. For our mobile data offloading system, however, the content dissemination delay becomes a secondary issue. In our investigation, we consider the delay as the system constraint by imposing the requirement of delivering each mobile data before its deadline  $T$ . Therefore, in the performance evaluation, we mainly focus on the average fraction of the offloaded data by the VDTN, which is the expected interest satisfaction clearly shown as the optimization goal in our problem (2). Although the mobile data for offloading do not have strict real-time requirements and can tolerate a certain amount of delay, the latency in the offloading does affect the QoS. It would be highly desired to simultaneously maximize the total amount of the offloaded data items and minimize the expected offloading delay. However, this is an open multiobjective optimization. Thus, in our simulation, we will evaluate the offloading efficiency of the proposed algorithms, by observing the results of interest satisfaction percentage, and also analyze how our proposed scheme allocates the system resource to mobile items according to the interests of subscribers.

We compared the performance of our scheme, labeled as **Coding Optimization (CO)**, which considers both optimal coding and optimal resource allocation, with the following schemes:

- 1) **Coding and Random Buffer Allocation (CR)**, in which all mobile data are coded by the minimum length  $s_p = 1$ , and each helper chooses the coded packets uniformly and randomly to fill its buffer until no more packets can be stored.
- 2) **No-coding and Greedy Buffer Allocation (NG)** [27], [28], where the system allocates the buffer for the uncoded data by a heuristic-based greedy algorithm. This algorithm represents the most updated work in the area of delay tolerant networking offloading [27].
- 3) **No-coding and Random Buffer Allocation (NR)** [27], which does not code data, and each helper chooses the data items randomly to fill its buffer until no more item can be stored.

### A. Simulation Setup

1) *Vehicular Mobility Trace:* Our evaluation was conducted on two real vehicular mobility traces, i.e., *Shanghai* [24] and *Beijing*, which recorded the positions of vehicles carrying GPS devices. Specifically, about 2100 operational taxis were running for the whole month of February 2007 in Shanghai city to collect *Shanghai* trace [24]. In collecting *Beijing* trace, we used mobility track logs obtained from 27 000 participating Beijing taxis carrying GPS receivers during the whole month of May in 2010. More specifically, we utilized the GPS devices to collect the taxi locations and timestamps and GPRS modules to report

the records every minute for moving taxis. *Beijing* trace is the largest vehicular data trace available.

In these vehicular traces, there were large number of nodes that rarely met with others. Therefore, we only selected nodes that had frequently contacted with others as simulation nodes. Specifically, for each of these two traces, we selected 150 nodes by ranking their contact times. As stated before, the mobility of the helpers has much more significant influence on the mobile data offloading system, and the nodes with more contact will be selected as the helpers by the service provided to enhance the system performance. Therefore, among these nodes, the most contacted 30 nodes were set as helpers by ranking their contact times, and the rest were used as subscribers. Moreover, the reason for setting the number of helpers to be 30 was to make sure the simulation results were robust for comparison and system parameter's variation. In all the experiments, the mobile data source, i.e., the service provider, obtained the average contact rates of every helper based on the contact counts from the beginning of the network simulation. We used the first half of the trace as this warm-up period for the mobile data source to accumulate necessary network information. Then, the mobile data were generated, coded, and stored in the buffers of all helpers and offloaded during the second half of the trace. After the simulation of the entire trace, we collected the system performance results.

2) *Mobile Data and Buffer*: In the simulation, there were  $C = 15$  mobile data items, and the sizes of data were generated randomly and uniformly in the range of [10 MB, 20 MB], while the data lifetime was set to 20%, 40%, 60%, 80%, and 100% of the simulated trace duration. We used the well-developed Tornado Z code for erasure coding, which is a class of erasure codes that have extremely fast encoding and decoding algorithms, proposed in [26]. Since Tornado Z code has an average decoding inefficiency of 1.054, we set  $\epsilon = 1.054$  for the erasure coding. The helpers' buffer sizes were randomly and uniformly generated in  $[0, 2l_a \text{ MB}]$ , where  $l_a$  was the average buffer size. In the simulation, the average buffer size was uniformly and randomly generated in [20 MB, 40 MB]. In order to define the subscribers' interests, we set the number of keywords to  $K = 20$  and assumed that the keywords  $k_1$  to  $k_{20}$  were ranked by their popularity. For each data item  $c_i$ ,  $i \in \{1, 2, \dots, 15\}$ , its describing subset of keywords was given by  $\mathcal{K}_{c_i} = \{k_i, k_{i+1}, \dots, k_{i+4}\}$ , with equal weight  $v_{k_i} = 1/5$  for each keyword. We used the normal distribution to generate the interest profile  $\mathcal{P}_s$  for user  $s$ . For keyword  $k_i$ , we assumed that the average interest of all the subscribers was  $I_i$  and used three distributions to generate  $I_i$  in order to obtain different interest distributions listed as follows.

- Exponential distribution:  $I_i = (e^{-i} / \sum_{j=1}^{20} e^{-j})$ . In this case, most of the subscribers' interests concentrated on the popular data.
- Zipf distribution with exponent 2:  $I_i = (1/i^2) / (\sum_{j=1}^{20} 1/j^2)$ . In this case, most of the subscribers' interests also concentrated on the popular data, but the difference between high and low popular data was smaller than the case of exponential distribution.

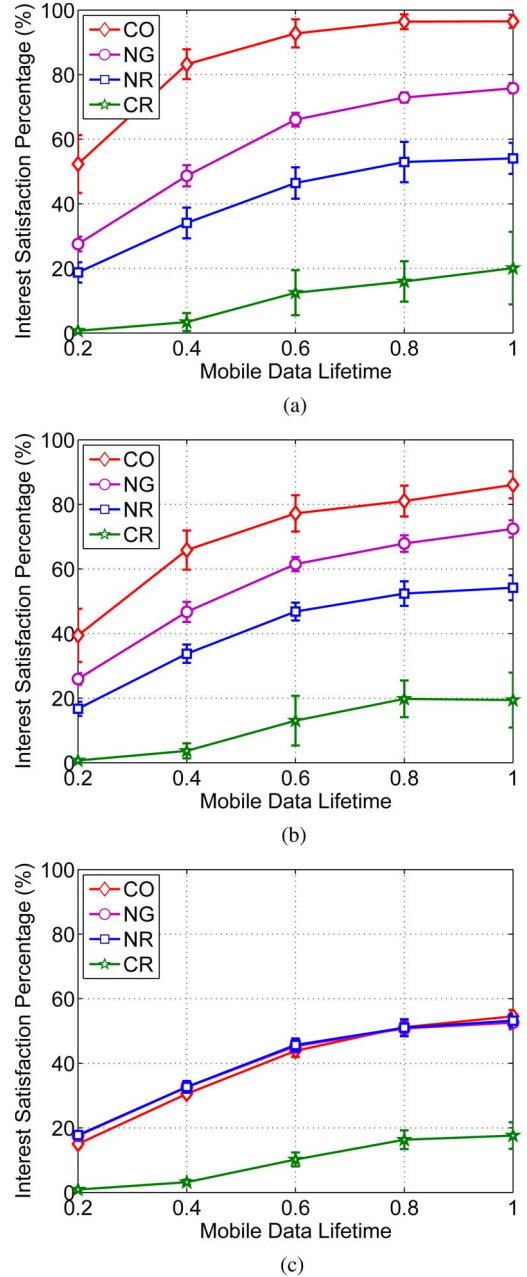


Fig. 4. Performance comparison of different schemes for *Beijing* trace with the fixed average buffer length of 30 MB and the variable average mobile data lifetime. (a) Exponential interests. (b) Zipf interests. (c) Uniform interests.

- Uniform distribution:  $I_i = 1/20$ . In this case, the subscribers' interests were uniformly distributed among the data.

## B. Offloading Efficiency

1) *Different Data Lifetime*: The results of subscribers' interest satisfaction for *Beijing* trace obtained by the different schemes with different interest distributions are shown in Fig. 4, where the simulation results were averaged over 30 runs by simulating the system with the buffer allocation strategy of each algorithm. As expected, as the mobile data lifetime increased, the interest satisfaction percentage increased since more packets were disseminated to the subscribers for longer data



lifetime. Specifically, Fig. 4(a) presents the results for the case of exponentially distributed interests, where it can be seen that our CO algorithm performed better than the three benchmarks, i.e., NG, NR, and CR algorithms, by about 53.2%, 1.3 times, and 19.1 times, respectively. It is worth noting that, when erasure coding was applied, the random allocation algorithm (CR) actually performed much worse than the noncoding random allocation algorithm (NR). Blindly applying erasure coding to all data may not be a good idea, as we have emphasized in the introduction section, and coding or not must be intelligently addressed, as our proposed scheme does. The efficiency of our optimal buffer allocation and coding design is clearly demonstrated by the results in Fig. 4(a). Similarly, the results under the case of Zipf distributed interests, as depicted in Fig. 4(b), show that our CO algorithm achieved the best performance, and it outperformed the NG, NR, and CR schemes by about 32.4%, 81.9%, and 13.7 times, respectively. Compared with the case of exponential interests shown in Fig. 4(a), the enhancement of our solution over the other algorithms was smaller. This is because Zipf distributed interests are more similar compared with exponentially distributed interests. Indeed, the results obtained with the case of uniformly distributed interests, as shown in Fig. 4(c), indicate that our CO algorithm achieved almost identical performance as the NG and NR did. It is seen that, with no variation in interests, the performance gain of our CO algorithm over these two benchmarks diminishes. We can also infer that our algorithm works more efficiently in the case of heterogeneous interest distribution. The underlying reason will be further discussed in Section V-C.

The results obtained for *Shanghai* trace are illustrated in Fig. 5, where again the performance were averaged over 30 runs. Similar observations to those for *Beijing* trace can be drawn. In particular, our proposed CO algorithm outperformed the NG and NR schemes by about 49.9% and 1.3 times, respectively, under the exponential distribution of interests, as shown in Fig. 5(a), whereas the performance gains of the CO algorithm over the NG and NR were about 30.4% and 79.7%, respectively, under the Zipf distribution of interests, as can be observed in Fig. 5(b). With no variation in interests, the three algorithms, i.e., the proposed CO and the benchmarks NG and NR, achieved similar performance, as shown in Fig. 5(c).

In our proposed CO scheme, Algorithm 1 provides the optimal resource allocation and coding policy, whereas Algorithm 2 only offers a heuristic solution to replicate the result obtained by Algorithm 1 into the helpers' buffers. This is because what critically influences the system performance is the resource allocation and coding policy, and a heuristic data replication algorithm is sufficient for replicating the optimal resource allocation and coding policy to helpers. We conducted a simulation to further verify this assertion using both *Shanghai* and *Beijing* traces with the three given interest distributions. Specifically, we ran the simulation 20 times to observe the system performance of interest satisfaction percentage, and in each run, we allocated the buffer by uniformly and randomly choosing the mobile data and helpers based on the solution obtained by Algorithm 1. The results so obtained (labeled as random replication) are shown in Table I, in comparison with the results obtained by Algorithm 2. It can be seen that the performance of

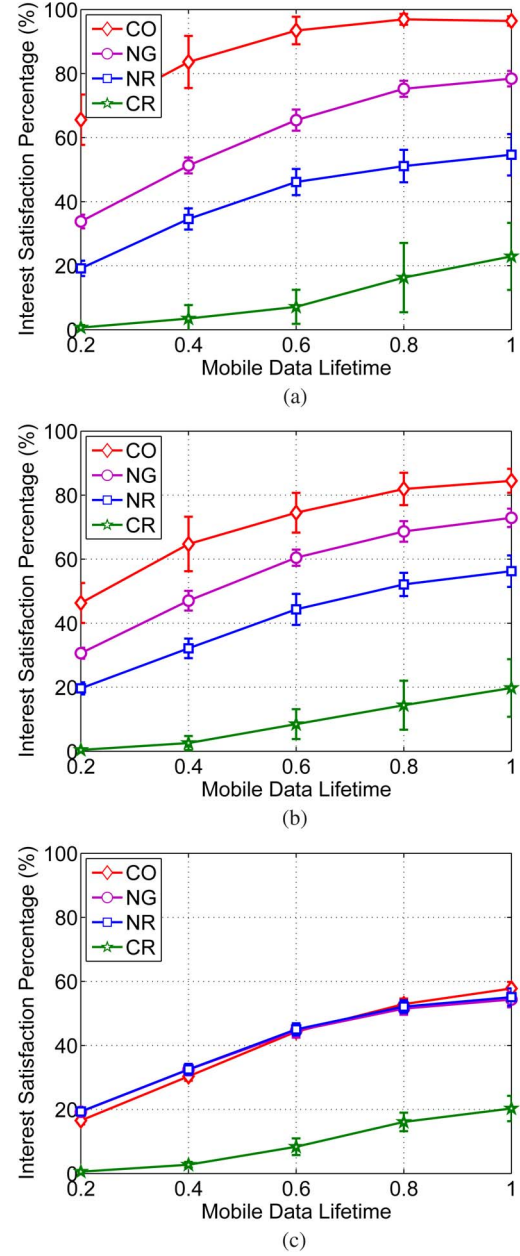


Fig. 5. Performance comparison of different schemes for *Shanghai* trace with the fixed average buffer length of 30 MB and the variable average mobile data lifetime. (a) Exponential interests. (b) Zipf interests. (c) Uniform interests.

Algorithm 2 is only marginally better than that of the random replication.

2) *Different Buffer Size*: Another important parameter that influences the overall system performance is the buffer size of helpers. We varied the average buffer size and simulated different algorithms to evaluate their performance. We only used the exponential and uniform distributions of interests since the performance that is achievable by the Zipf distribution is between these two cases. Figs. 6 and 7 present the results under *Beijing* and *Shanghai* traces, respectively. Similar observations can be drawn as the case of varying the average mobile data lifetime. Specifically, with the exponential distribution of interests, our CO algorithm outperformed the other coding- and non-coding-based schemes significantly, whereas with the uniform



TABLE I  
AVERAGE INTEREST SATISFACTIONS (%) AND STANDARD DEVIATIONS OBTAINED BY THE RANDOM DATA REPLICATION AND ALGORITHM 2 FOR SHANGHAI (SH) AND BEIJING (BJ) TRACES WITH THE FIXED AVERAGE BUFFER LENGTH OF 30 MB AND AVERAGE MOBILE DATA LIFETIME OF 100% DURATION OF THE TRACE, WHERE THE OPTIMAL RESOURCE ALLOCATION AND CODING POLICY ARE OBTAINED BY ALGORITHM 1 AND THE RESULTS ARE AVERAGED OVER 20 RANDOM RUNS

	random replication (SH)	Algorithm 2 (SH)	random replication (BJ)	Algorithm 2 (BJ)
Exponential interests	$96.53 \pm 0.487$	$97.24 \pm 0.250$	$96.27 \pm 0.559$	$97.16 \pm 0.265$
Zipf interests	$84.01 \pm 0.582$	$85.42 \pm 0.459$	$82.96 \pm 0.694$	$84.35 \pm 0.489$
Uniform interests	$58.37 \pm 0.545$	$61.61 \pm 0.274$	$55.77 \pm 0.552$	$58.73 \pm 0.242$

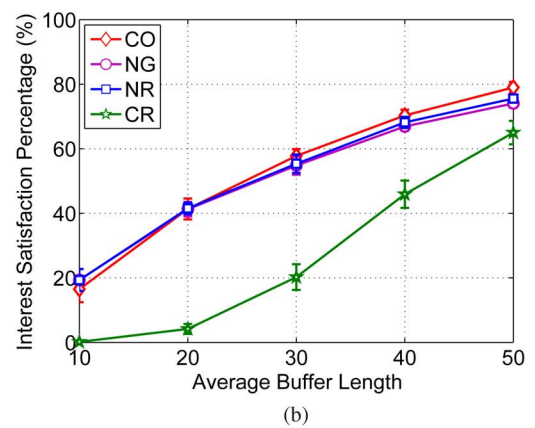
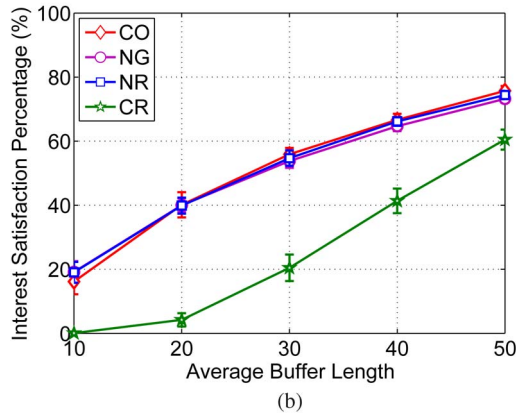
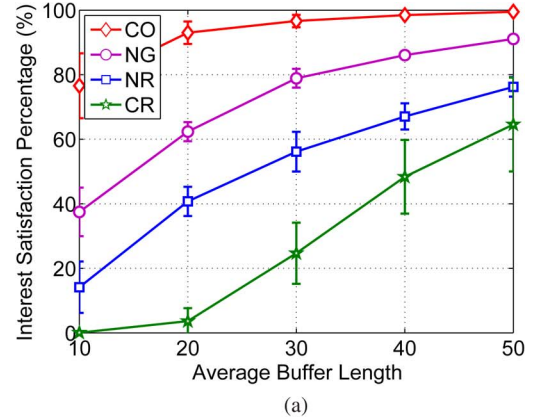
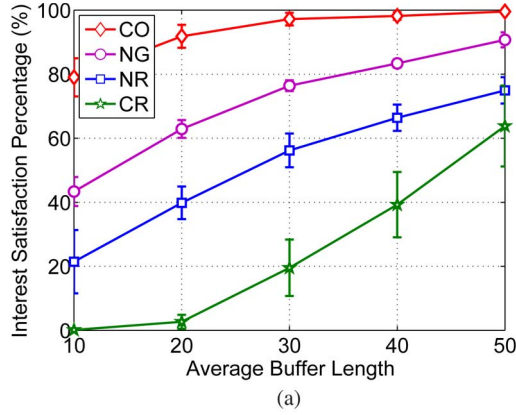


Fig. 6. Performance comparison of different schemes for *Beijing* trace with the fixed average mobile data lifetime of 100% duration of the trace and the variable average buffer length. (a) Exponential interests. (b) Uniform interests.

Fig. 7. Performance comparison of different schemes for *Shanghai* trace with the fixed average mobile data lifetime of 100% duration of the trace and the variable average buffer length. (a) Exponential interests. (b) Uniform interests.

distribution of interests, our proposed CO algorithm and the existing NR and NG schemes achieved similar performance. These results again confirm the effectiveness of our proposed approach in the environment of heterogeneously distributed interests.

### C. Interest Distribution and Resource Allocation

As previously shown, the performance gain of our proposed CO scheme over other existing schemes diminishes when there exists no variation in the interests to mobile data items. It is therefore intriguing to investigate how our optimal Algorithm 1 allocates the system resource to mobile items according to the interests of subscribers. In the simulation investigation, we restricted to the two types of mobile data with the lengths of 10 and 20 MB, respectively, as well as considered the exponential and uniform interest distributions. The results obtained for *Beijing* and *Shanghai* traces are shown in Figs. 8 and 9, respectively, where the  $x$ -axis represents the total interest of each

mobile data  $W_c$ , and the  $y$ -axis represents the network resource  $\eta_c$ , which is allocated by Algorithm 1, to each mobile data. In Figs. 8(a) and 9(a), which depict the results obtained under the exponentially distributed interests, it is shown that the interests on different data vary approximately from 0 to 8, which show that, as expected, under the exponential interest distribution, some data are interested by a large number of subscribers, but some other items may not be interesting to most subscribers. In terms of resource allocation, we can see that the allocated network resource to each data is approximately proportional to the total interest. Therefore, the random resource allocation policy will perform badly under this environment. It can also be seen that, with regard to the coding policy, our proposed CO algorithm encodes the data that have larger interests, whereas the data with small interests will be assigned with less resource or no resource without using coding. Thus, our algorithm allocates more resource to the data that are interested by a larger number of subscribers, and in this way, it achieves a better system performance. For the results obtained under the uniform interest

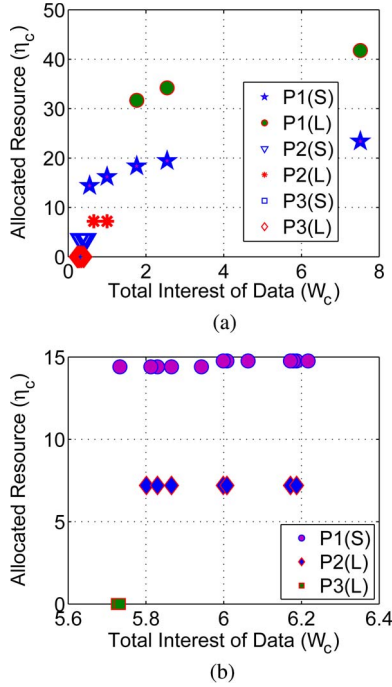


Fig. 8. How Algorithm 1 allocates resource to mobile data according to total interests of all the subscribers for *Beijing* trace with the average mobile data lifetime of 100% duration of the trace and the average mobile data length of 10 MB (S) and 20 MB (L), where P1 denotes the coded data, P2 denotes the uncoded data, and P3 denotes the data that have no resource assigned to. (a) Exponential interests. (b) Uniform interests.

distribution shown in Figs. 8(b) and 9(b), all the users have almost the same interests that only vary from 5.6 to 6.4. It can be seen that Algorithm 1 encodes all the small-size data items and allocates an equal-high resource to them and that it does not encode the large-size data items and allocate an equal resource to them with the exception of a few relatively low-interest items that get no resource. Under the uniform interest distribution, the network resource can be allocated more equally, and this is the reason why the performance gain of our proposed CO algorithm over the NG and NR algorithms diminishes.

## VI. CONCLUSION

We have studied the problem of optimal mobile data offloading in opportunistic vehicular networks in a realistic environment, where the network is heterogeneous in terms of the disseminated data being multitypes with different delay sensitivities and lengths, as well as the network helpers' storages being limited with different buffer sizes. By formulating this challenging problem as a utility function maximization problem, we have designed an efficient scheme to decide the optimal coding policy and to allocate the helpers' buffer optimally. Extensive simulation results have demonstrated that, for the realistic environment where subscribers have different interests to different mobile data items, our algorithm achieves significantly better performance than the existing algorithms that are traditionally used to solve this type of challenging problems.

In this paper, we have replaced the individual contact rate of a helper and a subscriber by the average contact rate of

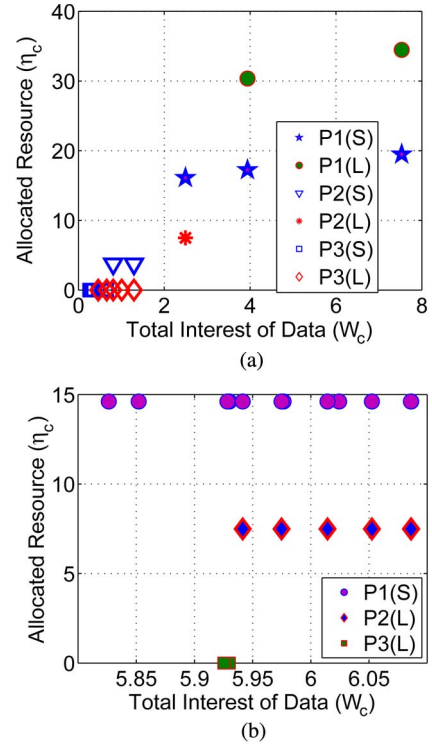


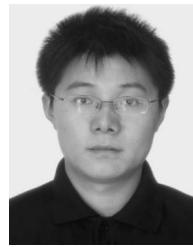
Fig. 9. How Algorithm 1 allocates resource to mobile data according to total interests of all the subscribers for *Shanghai* trace with the average mobile data lifetime of 100% duration of the trace and the average mobile data length of 10 MB (S) and 20 MB (L), where P1 denotes the coded data, P2 denotes the uncoded data, and P3 denotes the data that have no resource assigned to. (a) Exponential interests. (b) Uniform interests.

a helper to all the subscribers, in order to make analysis and design tractable. Our future work will investigate whether it is feasible to relax this assumption in the formulation of the system expected interest satisfaction optimization and in the derivation of the optimal solution.

## REFERENCES

- [1] M. Khabazian, S. Aissa, and M. Mehmet-Ali, "Performance modeling of message dissemination in vehicular ad hoc networks with priority," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 61–71, Jan. 2011.
- [2] Cisco visual networking index: Global mobile data traffic forecast update, 2011–2016, Cisco Syst., Inc., San Jose, CA, USA. [Online]. Available: [http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white\\_paper\\_c11-520862.html](http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html)
- [3] K. Lee, I. Rhee, J. Lee, Y. Yi, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 425–426, Apr. 2010.
- [4] J. Zhao and G. Cao, "VADD: Vehicle-assisted data delivery in vehicular ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 3, pp. 1910–1922, May 2008.
- [5] D. Câmara, N. Frangiadakis, F. Filali, and C. Bonnet, "Vehicular delay tolerant networks," in *Handbook of Research on Mobility and Computing: Evolving Technologies and Ubiquitous Impacts*, M. M. Cruz-Cunha and F. Moreira, Eds. Hershey, PA, USA: IGI Global, 2011, pp. 356–367.
- [6] J. Burgess, B. Gallagher, D. Jensen, and B. Levine, "MaxProp: Routing for vehicle-based disruption-tolerant networks," in *Proc. 25th IEEE INFOCOM*, Barcelona, Spain, Apr. 23–29, 2006, pp. 1–11.
- [7] V. N. G. J. Soares, F. Farahmand, and J. Rodrigues, "A layered architecture for vehicular delay-tolerant networks," in *Proc. IEEE Symp. Comput. Commun.*, Sousse, Tunisia, Jul. 5–8, 2009, pp. 122–127.
- [8] J. Widmer and J. Le Boudec, "Network coding for efficient communication in extreme networks," in *Proc. ACM SIGCOMM Workshop Delay-Tolerant Netw.*, Philadelphia, PA, USA, Aug. 22–26, 2005, pp. 284–291.

- [9] S. Jain, M. Demmer, R. Patra, and K. Fall, "Using redundancy to cope with failures in a delay tolerant network," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 4, pp. 109–120, Oct. 2005.
- [10] Y. Liao, K. Tan, Z. Zhang, and L. Gao, "Estimation based erasure-coding routing in delay tolerant networks," in *Proc. Int. Conf. Wireless Commun. Mobile Comput.*, Vancouver, Canada, Jul. 3–6, 2006, pp. 557–562.
- [11] L. Hu, J.-Y. Le Boudec, and M. Vojnovic, "Optimal channel choice for collaborative ad-hoc dissemination," in *Proc. 29th IEEE INFOCOM*, San Diego, CA, USA, Mar. 14–19, 2010, pp. 1–9.
- [12] B. Han, P. Hui, V. S. Anil Kumar, M. V. Marathe, G. Pei, and A. Srinivasan, "Cellular traffic offloading through opportunistic communications: A case study," in *Proc. 5th ACM Workshop Challenged Netw.*, Chicago, IL, USA, Sep. 20–24, 2010, pp. 31–38.
- [13] J. Reich and A. Chaintreau, "The age of impatience: Optimal replication schemes for opportunistic networks," in *Proc. 5th Int. Conf. Emerging Netw. Exp. Technol.*, Rome, Italy, Dec. 1–4, 2009, pp. 85–96.
- [14] R. Bhatia, G. Narlikar, I. Rimac, and A. Beck, "UNAP: User-centric network-aware push for mobile content delivery," in *Proc. 28th IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 19–25, 2009, pp. 2034–2042.
- [15] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proc. 8th Int. Conf. Mobile Syst., Appl. Services*, San Francisco, CA, USA, Jun. 15–18, 2010, pp. 209–222.
- [16] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, and M. D. De Amorim, "Relieving the wireless infrastructure: When opportunistic networks meet guaranteed delays," *Arxiv preprint arXiv:1007.5459*, 2010.
- [17] B. D. Higgins, A. Reda, T. Alperovich, J. Flinn, T. J. Giuli, B. Noble, and D. Watson, "Intentional networking: Opportunistic exploitation of mobile network diversity," in *Proc. 16th ACM Int. Conf. Mobile Comput. Netw.*, Chicago, IL, USA, Sep. 20–24, 2010, pp. 73–84.
- [18] U. Shevade, Y.-C. Chen, L. Qiu, Y. Zhang, V. Chandar, M. K. Han, H. H. Song, and Y. Seung, "Enabling high-bandwidth vehicular content distribution," in *Proc. 6th Int. Conf. Emerging Networking Experiments and Technologies*, Philadelphia, PA, USA, Nov. 30–Dec. 3, 2010, pp. 12–23.
- [19] P. Costa, C. Mascolo, M. Musolesi, and G. P. Picco, "Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 5, pp. 748–760, Jun. 2008.
- [20] S. Ioannidis, A. Chaintreau, and L. Massoulié, "Optimal and scalable distribution of content updates over a mobile social network," in *Proc. 28th IEEE INFOCOM*, Rio De Janeiro, Brazil, Apr. 19–25, 2009, pp. 1422–1430.
- [21] H. Luo, X. Meng, R. Ramjee, P. Sinha, and L. Li, "The design and evaluation of unified cellular and ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 6, no. 9, pp. 1060–1074, Sep. 2007.
- [22] B. Han, P. Hui, and A. Srinivasan, "Mobile data offloading in metropolitan area networks," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 14, no. 4, pp. 28–30, Oct. 2011.
- [23] K. Lee, Y. Yi, J. Jeong, H. Won, I. Rhee, and S. Chong, "Max-contribution: On optimal resource allocation in delay tolerant networks," in *Proc. 29th IEEE INFOCOM*, San Diego, CA, Mar. 14–19, 2010, pp. 1–9.
- [24] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li, and L. Ni, "Recognizing exponential inter-contact time in VANETs," in *Proc. 29th IEEE INFOCOM*, San Diego, CA, USA, Mar. 14–19, 2010, pp. 1–5.
- [25] C. Fragouli, J. Widmer, and J. Y. Le Boudec, "On the benefits of network coding for wireless applications," in *Proc. 4th Int. Symp. Modeling Optim. Mobile, Ad Hoc Wireless Netw.*, Boston, MA, USA, Apr. 3–6, 2006, pp. 1–6.
- [26] J. Byers, M. Luby, and M. Mitzenmacher, "A digital fountain approach to asynchronous reliable multicast," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 8, pp. 1528–1540, Oct. 2002.
- [27] Y. Li, G. Su, P. Hui, D. Jin, L. Su, and L. Zeng, "Multiple mobile data offloading through delay tolerant networks," in *Proc. 6th ACM Workshop Challenged Netw.*, Las Vegas, NV, USA, Sep. 19–23, 2011, pp. 43–48.
- [28] A. Kulik, H. Shachnai, and T. Tamir, "Maximizing submodular set functions subject to multiple linear constraints," in *Proc. 20th Annu. ACM-SIAM Symp. Discrete Algorithms*, New York, Jan. 4–6, 2009, pp. 545–554.
- [29] S. Jain, M. Demmer, R. Patra, and K. Fall, "Using redundancy to cope with failures in a delay tolerant network," in *Proc. of ACM SIGCOMM CHANTS*, Philadelphia, PA, USA, Aug. 22–26, 2005, pp. 109–120.
- [30] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Medard, and J. Crowcroft, "XORs in the air: Practical wireless network coding," in *Proc. ACM SIGCOMM CHANTS*, Pisa, Italy, Sep. 11–15, 2006, pp. 243–254.
- [31] Y. Lin, B. Li, and B. Liang, "Efficient network coded data transmissions in disruption tolerant networks," in *Proc. 27th IEEE INFOCOM*, Phoenix, AZ, USA, Apr. 15–17, 2008, pp. 2180–2188.
- [32] U. Lee, J. Park, J. Yeh, G. Pau, and M. Gerla, "Code torrent: Content distribution using network coding in VANET," in *Proc. ACM 1st Int. Workshop Decentralized Resource Sharing Mobile Comput. Netw.*, Los Angeles, CA, USA, Sep. 29–29, 2006, pp. 1–5.
- [33] M. Sathiamoorthy, A. G. Dimakis, B. Krishnamachari, and F. Bai, "Distributed storage codes reduce latency in vehicular networks," in *Proc. IEEE INFOCOM*, 2012, pp. 2646–2650.
- [34] S. Lee, U. Lee, K. Lee, and M. Gerla, "Content distribution in VANETs using network coding: The effect of disk I/O and processing O/H," in *Proc. IEEE 5th SECON*, San Francisco, CA, USA, Jun. 16–20, 2008, pp. 117–125.
- [35] M. Li, Z. Yang, and W. Lou, "CodeOn: Cooperative popular content distribution for vehicular networks using symbol level network coding," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 223–235, Jan. 2011.
- [36] N. Abdullah, A. Doufexi, and R. Piechocki, "Raptor codes for infrastructure-to-vehicular broadcast services," in *Proc. IEEE VTC Fall Conf.*, San Francisco, CA, USA, Sep. 5–8, 2011, pp. 1–5.



**Yong Li** (M'09) received the B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007 and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012.

He is a Postdoctoral Researcher with Tsinghua University. His research interests include mobile delay-tolerant networks, including forwarding policy design, buffer management design, and performance evaluation; mobility modeling; and mobility management in next-generation wireless IP networks, including Mobile IP, Session Initiation Protocol, Proxy Mobile IP, cross-layer design, multicast mobility, modeling for mobility performance evaluation, enhancing handoff performance, and proposing mobility management architecture.

Dr. Li serves as a Paper Reviewer for the IEEE International Conference on Communications; the Vehicular Technology Conference; the International Conference on Information Networking; the International Symposium on Personal, Indoor, and Mobile Radio Communications; the Asia-Pacific Conference on Communications; and many others.



**Depeng Jin** (M'09) received the B.S. and Ph.D. degrees from Tsinghua University, Beijing, China, in 1995 and 1999, respectively, both in electronics engineering.

He is an Associate Professor with Tsinghua University, where he is the Vice Chair of the Department of Electronic Engineering. His research interests include telecommunications, high-speed networks, application-specific integrated circuit design, and future Internet architecture.

Dr. Jin was a recipient of the National Scientific and Technological Innovation Prize (Second Class) in 2002.



**Zhaocheng Wang** (M'09–SM'11) received the B.S., M.S., and Ph.D. degrees from Tsinghua University, Beijing, China, in 1991, 1993, and 1996, respectively.

From 1996 to 1997, he was with Nanyang Technological University, Singapore, as a Postdoctoral Fellow. From 1997 to 1999, he was with OKI Techno Centre Pte. Ltd., Singapore, first as a Research Engineer and then as a Senior Engineer. From 1999 to 2009, he was with Sony Deutschland GmbH, Cologne, Germany, first as a Senior Engineer and then as a Principal Engineer. He is currently a Professor with the Department of Electronic Engineering, Tsinghua University. His research interests include wireless communications, digital broadcasting, and millimeter-wave communications. He is the holder of 22 granted U.S. and EU patents and is the author of over 60 technical papers.

Dr. Wang is a Fellow of the Institution of Engineering and Technology. He served as a Technical Program Committee Co-Chair/Member of many international conferences.



**Lieguang Zeng** received the B.S. degree from Tsinghua University, Beijing, China, in 1970.

Since 1970, he has been with the Department of Electronic Engineering, Tsinghua University, where he is currently a Professor. He is the author of more than 150 research papers and is the holder of four patents. His research interests include telecommunications, high-speed networks, application-specific integrated circuit design, and future Internet architecture.

Prof. Zeng was a recipient of the National Scientific and Technological Innovation Prize (Second Class) three times (in 1987, in 1991, and in 2002) and the Provincial and Ministerial Level Science and Technology Progress Prize eight times.



**Sheng Chen** (M'90–SM'97–F'08) received the B.Eng. degree in control engineering from the East China Petroleum Institute, Dongying, China, in January 1982, the Ph.D. degree in control engineering from the City University, London, U.K., in September 1986, and the D.Sc. degree from University of Southampton, Southampton, U.K., in 2005.

From 1986 to 1999, he held research and academic appointments with the Universities of Sheffield, Edinburgh, and Portsmouth, all in the U.K. Since 1999, he has been with the Department of Electronics and Computer Science, University of Southampton, where he is currently a Professor in intelligent systems and signal processing. He is a Distinguished Adjunct Professor with King Abdulaziz University, Jeddah, Saudi Arabia. He is the author of over 480 research papers. He is an Institute for Science Information Highly Cited Researcher in the engineering category (March 2004). His recent research interests include adaptive signal processing, wireless communications, modeling and identification of nonlinear systems, neural network and machine learning, intelligent control system design, evolutionary computation methods, and optimization.

Dr. Chen is a Chartered Engineer and a Fellow of the Institution of Engineering and Technology.